DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 12 | Issue 1 Article 19

5-1-2013

The Probit Link Function in Generalized Linear Models for Data Mining Applications

Mehdi Razzaghi
Bloomsburg University, Bloomsburg, PA

Journal of Modern Applied Statistical Methods
May 2013, Vol. 12, No. 1, 164-169

1538 – 9472/13/$95.00

The Probit Link Function in Generalized Linear Models for Data Mining Applications

Mehdi Razzaghi
Bloomsburg University,
Bloomsburg, PA

The use of logistic regression for outcome classification of dichotomous variables is well known in data mining applications. The estimated probability of the logit transformation belongs to the class of canonical link functions that follow from particular probability distribution functions. A closely related model is the probit link which can be used for binary responses. Although the probit link is not canonical, in some cases the overall fit of the model can be improved by using non-canonical link functions. This article reviews the properties of the probit link function and discusses its applications in data mining problems. Contrasts and comparisons are made with the logistic link function and an example provides further illustration.

Key words: Probit, logistic, linear models, data mining, link functions.

Introduction

The problem of outcome classification of qualitative data is a major task in data mining. The goal of classification is to accurately predict the target class for each case in the data. Specifically, in binary classification, the target attribute has only two possible outcomes and fast and accurate classifiers are highly desirable. Several predictive models such as naïve Bayes, classification trees, support vector machine and k-nearest neighbor have traditionally been used with some success. However, recently the use of logistic regression has found more widespread popularity and the method has attracted the attention of several practitioners. The advantage of such a model is that it transforms information about the binary dependent variable into an unbounded continuous variable and estimates a regular multivariate model. Komarek and Moore (2005) present a simple parameter-free implementation of the logistic regression and demonstrate that the model is sufficiently fast and accurate for classification of binary outcomes in large real-world datasets. Maalouf (2011) presented an overview of various aspects of logistic regression, calling it one of the most important and one of the most widely used data mining techniques. Comparatively less attention has been focused on a similar, but slightly different probit model. The difference between the two models is that the logistic model is based on the logit transformation while the probit model uses the inverse Gaussian link. In most cases, the classification outcome is similar for the two models even though the underlying distributions are different.

Mehdi Razzaghi is a Professor in the Department of Mathematics. He holds a Ph.D. in statistics from the University of London, England and has over 30 years of teaching and research experience. His primary research interests are in distribution theory and application of statistics in environmental sciences. Email him at: mrazzagh@bloomu.edu.

Generalized Linear Models

In linear regression analysis there is a random component $\boldsymbol{Y}$ which identifies the response variable and several explanatory variables (features or attributes) $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, …, $\boldsymbol{X_p}$. The response variable is expressed as a linear predictor of the explanatory variables,

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \cdots + \beta_p \boldsymbol{X_p} + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{\varepsilon}$ is the error term and $\boldsymbol{Y}$, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, … $\boldsymbol{X_p}$ and $\boldsymbol{\varepsilon}$ are all n × 1 vectors where n is the number of instances or the sample size. In matrix form, the model can be expressed as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ and $\boldsymbol{X}$ is the n × p data matrix often referred to as the design matrix. The assumption of linearity in this model can be too restrictive and in many cases is unrealistic. In addition, the model assumes that response variable $\boldsymbol{Y}$ has a normal distribution with a constant variance; that is, if $\boldsymbol{Y} = (y_1, y_2, \dots, y_n)^T$ then it is assumed that $E(\boldsymbol{Y}) = \boldsymbol{\mu} = \boldsymbol{X\beta}$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)^{\boldsymbol{T}}$ and that y_i has a normal distribution with mean $\mu_i = E(y_i)$ and variance σ^2 for i = 1, 2, …, n. These assumptions, in many instances, may not be justified. To overcome this problem, the assumptions are relaxed by allowing $\boldsymbol{\mu}$ and $\boldsymbol{X\beta}$ to be related by a so-called link function, g, so that

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta}. \quad (3)$$

In this way, the normal model becomes a special case of generalized linear model in which the link function is the identity function. Therefore, in general, for distributions other than normal that fit the data, a suitable link function can be determined. Specifically, if responses are binary as in the binomial distribution, the two popular link functions are the logit transformation, $\log\left(\frac{\mu_i}{1-\mu_i}\right)$ and the probit transformation, $\Phi^{-1}(\mu_i)$ where Φ is the cumulative distribution function of the standard normal distribution. In the case of logit transformation, the outcome probability is assumed to have the logistic distribution, whereas in the case of the probit link, the distribution of the outcome probability does not have an easily interpretable form. This is why the logit transformation belongs to the canonical family of link functions while the probit link is not canonical.

Logistic Regression Model

Assume that response variable $\boldsymbol{Y}$ is binary and let $P(y_i = 1) = \pi_i$ be the success probability for the i^{th} measurement. Then, it can be shown that $E(y_i) = \pi_i$ and $V(y_i) = \pi_i(1 - \pi_i)$. In ordinary regression, $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots \pi_n)^T$ is modeled as a linear function $\boldsymbol{X}$ with a constant variance. However, because the expected value and variance of the response variable are not constant, ordinary linear regression does not apply. Moreover, it can be shown that the relationship between $\boldsymbol{\pi}$ and $\boldsymbol{X}$ is generally not linear. Constant changes in the explanatory variables usually have less impact in the success probability π_i especially when π_i is close to 0 or 1. Sigmoid shaped curves are often more realistic for such relationships, and among these, the most commonly used is the logistic function defined as

$$\pi_i = (1 + e^{-\boldsymbol{x_i\beta}})^{-1} \quad (4)$$

where $\boldsymbol{x_i}$ is the i^{th} row of the $\boldsymbol{X}$ matrix, that is, the i^{th} record in the dataset. Taking the logarithm of the odds ratio, called the logit transformation, from equation (4) results in the logistic regression model

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x_i\beta} + \epsilon_i \quad (5)$$

where ϵ_i is the error term. A vast body of literature exists regarding methods for fitting a logistic regression model. The popular maximum likelihood method is used to estimate model parameters. Note that if the responses are independent, then by applying the Bernoulli distribution, the likelihood for n binary observations as a function of the parameters is

$$L(\boldsymbol{\beta}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \quad (6)$$

and the log-likelihood is given by

$$log\, L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \log \pi_i + (1 - y_i) \log (1 - \pi_i)) \quad (7)$$

which must be maximized in order to derive the maximum likelihood estimates of the model

parameters. The first derivative of equation (7) is referred to as the score function (McCullagh & Nelder, 1989). Because no analytic solution for the maximum likelihood estimates can be derived, numerical methods are used. The most popular method for numerical derivation of the parameter estimates is an adaptation of the Newton-Raphson method, called the Iteratively Re-Weighted Least Squares (IRLS). In this method, a new set of weights are estimated at each iteration (Hosmer & Lemeshow, 2000; Hilbe, 2009; Maalouf, 2011).

Probit Regression Model

The probit model is another sigmoid-shaped curve used in modeling dichotomous outcome variables. For this model, the link function, called the probit link, uses the inverse of the cumulative distribution function of the standard normal distribution to transform probabilities to the standard normal variable. Thus

$$\Phi^{-1}(\pi_i) = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \qquad (6)$$

where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt. \qquad (7)$$

The use of the probit regression model dates back to Bliss (1934). Bliss was interested in finding an effective pesticide to control insects that fed on grape leaves (Greenberg, 1980). He found that the relationship between response to a dose of pesticide was sigmoid and he applied the probit transformation to transform the sigmoid shape dose-response curve to a linear relationship. His ideas were later generalized in a book by Finney (1985) where the applications of probit analysis in toxicological experiments were explored. According to some sources, probit analysis remains the preferred method in understanding dose-response relationships. In data mining, however, this application remains fairly unknown even though most popular statistical software such as SPSS, SAS and R carry functions for probit regression.

The probit model has also found popularity in economics. Cramer (2003) provides a survey of the early origins of the model. In comparing the probit model to the logistic model, many authors believe that there is little theoretical justification in choosing one formulation over the other in most circumstances involving binary responses. The logit model is considered to be computationally simpler but it is based on a more restrictive assumption of error independence, although many other generalizations have dealt with that assumption as well. By contrast, the probit model assumes that random errors have a multivariate normal distribution. This assumption makes the probit model attractive because the normal distribution provides a good approximation to many other distributions. The model does not rely on the assumption of error independence and econometricians utilize a general random utility model to describe the correlation. Hausman and Wise (1978) defined the covariance probit model and used it in economic applications. The model parameters are estimated by using the generalized least squares

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} \qquad (7)$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the errors and $\boldsymbol{Z}$ is the vector of observed probits:

$$\boldsymbol{Z} = (\Phi^{-1}(\pi_1), \Phi^{-1}(\pi_2), \ldots, \Phi^{-1}(\pi_n))^{\boldsymbol{T}}.$$

In practice, $\boldsymbol{\Sigma}$ is replaced by some consistent estimate of the covariance matix. This makes the probit model computationally more complicated; however, as argued by Judge, et al. (1980) the probit model is more flexible. The use of noncanonical link functions is not prohibited by the fact that they are more computationally complex. Czado and Munk (2000) argued that, in some applications, the overall fit of the model as measured by the p-value of the goodness-of-fit statistics can be improved significantly by the use of a noncanonical link.

Applications in Data Mining

Now consider a data set with binary responses and consider analyzing the data using both logistic regression and probit regression models. This study used a data set regarding credit reliability of individuals. If a financial institution gives a loan based on a credit to a person, they clearly would be interested in

estimating the person's ability to pay the loan back, thus, the aim is to model credit reliability as a function of the person's risk factors (covariates). The data set consists of 1,000 customers from a bank in southern Germany. The response variable is in dichotomous form using 0 for a reliable client and 1 for not reliable. The data set, which is described in Fahrmeir and Hamerele (1994), consists of the following 20 covariates that were considered to be important in credit evaluation of a client:

X_1: Running account (trichotomous)

X_2: Duration of credit (month)

X_3: Payment of previous credits

X_4: Purpose of Credit

X_5: Amount of credit

X_6: Value of savings or stocks

X_7: Employment history

X_8: Credit payment as percentage of income

X_9: Marital status

X_{10}: Further debtors/guarantors

X_{11}: Number of years in current household

X_{12}: Most valuable assets

X_{13}: Age

X_{14}: Running credits in other institutions

X_{15}: Own/rent

X_{16}: Number of previous credits at this bank

X_{17}: Occupation

X_{18}: Number of persons entitled to maintenance

X_{19}: Telephone

X_{20}: Foreign or national worker

Fahrmeir and Hamerele (1994) used a logit model to analyze a subset of the data with seven covariates; later Fahrmeir and Tutz (1997) used the data set to illustrate an example in logistic regression. The data set was further analyzed in depth as a case study by Giudici (2003) where a descriptive analysis of the data set was also included. PROC LOGISTIC and PROC PROBIT in SAS were used to analyze the data. Table 1 provides the parameter estimates for each covariate under the two modeling structures together with standard error, the value of the Wald Chi square and the significance level as measured by the p-value. It can be observed that, although the values of the parameter estimates are different – as they should be, both models produce very similar results and point to the same set of parameters as significant. The standard error of the estimates appears to be smaller for some variables in the logistic model and larger in others. The predicted values and other standard statistics were computed and again very similar results were obtained under both models.

Discussion

In data mining, there is a strong urge to use logistic regression as one of the main approaches for classification of binary responses. Komarek and Moore (2005) presented an argument in introducing the logistic regression as a core in data mining tools. A large body of literature exists on the use of logistic regression in data mining applications. Comparatively less is known about a similar, but intrinsically different approach of probit regression. This article introduced this model as another powerful and useful approach for modeling binary data. Many authors have used the probit model in other applications with success; for example, Shariff, et al. (2009) compared the two models for estimating the strength of gear teeth. The goal herein was to present the probit regression to the data mining community; it was not to introduce the probit model as a rival to the logistic model, but rather as an alternative. Experience shows that in most situations the two approaches produce similar results although some differences exist. This similarity is not necessarily sustained when multivariate responses are used. Further research is needed to investigate the advantages or disadvantages in using one model over the other in data mining applications.

Table 1: Maximum Likelihood Estimate of Model Parameters

	Logistic Regression Parameters				Probit Regression Parameters			
	Estimate	SE	Chi-Square	Pr > ChiSq	Estimate	SE	Chi-Square	Pr > ChiSq
Intercept	3.9940	1.0238	15.2178	<.0001	2.2004	0.5729	14.7500	0.0001
X_1	-0.5799	0.0700	68.5630	<.0001	-0.3424	0.0400	73.4300	<.0001
X_2	0.0246	0.0087	7.9300	0.0049	0.0140	0.0051	7.4600	0.0063
X_3	-0.3822	0.0874	19.1204	<.0001	-0.2220	0.0501	19.6300	<.0001
X_4	-0.0315	0.0301	1.0980	0.2947	-0.0173	0.0173	1.0000	0.3180
X_5	0.0001	0.0000	5.4199	0.0199	0.0001	0.0000	5.7000	0.0170
X_6	-0.2391	0.0583	16.8395	<.0001	-0.1312	0.0326	16.2000	<.0001
X_7	-0.1517	0.0712	4.5444	0.0330	-0.0876	0.0413	4.4800	0.0342
X_8	0.2983	0.0828	12.9949	0.0003	0.1748	0.0478	13.3800	0.0003
X_9	-0.2574	0.1157	4.9473	0.0261	-0.1477	0.0668	4.8800	0.0271
X_{10}	-0.3473	0.1777	3.8188	0.0507	-0.1804	0.1011	3.1800	0.0744
X_{11}	0.0141	0.0774	0.0332	0.8553	0.0076	0.0453	0.0300	0.8660
X_{12}	0.1828	0.0910	4.0367	0.0445	0.1102	0.0528	4.3500	0.0369
X_{13}	-0.0089	0.0082	1.1807	0.2772	-0.0055	0.0047	1.3500	0.2457
X_{14}	-0.2419	0.1111	4.7428	0.0294	-0.1381	0.0650	4.5100	0.0337
X_{15}	-0.2931	0.1677	3.0542	0.0805	-0.1629	0.0981	2.7600	0.0969
X_{16}	0.2436	0.1610	2.2894	0.1303	0.1460	0.0921	2.5100	0.1130
X_{17}	-0.0189	0.1367	0.0191	0.8901	-0.0101	0.0805	0.0200	0.8996
X_{18}	0.1708	0.2319	0.5421	0.4616	0.1133	0.1348	0.7100	0.4007
X_{19}	-0.2947	0.1880	2.4567	0.1170	-0.1692	0.1089	2.4100	0.1204
X_{20}	-1.1582	0.6078	3.6317	0.0567	-0.6389	0.3292	3.7700	0.0523

References

Bliss, C. L. (1934). Methods of probits. *Science*, *79*, 38-39.

Cramer, J. S. (2003). *Logit models from economics and other fields*. Cambridge, MA: Cambridge University Press.

Czado, C., & Munk, A. (2000). Noncanonical links in generalized linear models: When is the effort justified? *Journal of Statistical Planning and Inference*, *87*, 317-345.

Fahrmeir, L., & Hamerele, A. (1984). *Multivariate statistische verfahren*. Berlin: Springer Verlag.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modeling based on generalized linear models*. Berlin: Springer Verlag.

Finney, D. J. (1971). *Probit analysis*. Cambridge, MA: Cambridge University Press.

Giudici, P. (2003). *Applied data mining, statistical methods for business and industry*. Chichester, MA: Wiley.

Greenberg, B. G. (1980). Chester I. Bliss, 1899-1979. *International Statistical Review*, *8*, 135-136.

Hausman, J. A., & Weis, D. A. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica*, *46*, 403-426.

Hilbe, J. M. (2009). *Logistic regression models*. Boca Raton, FL: Chapman & Hall/CRC Press.

Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression. New York, NY: Wiley.

Judge, G. G., Griffith, W. E., Hill, R. C., & Lee, T. C. (1985). The theory and practice of econometrics. New York, NY: Wiley.

Komarek, P., & Moore, A. (2005). *Making logistic regression a core data mining tool: A practical investigation of accuracy, speed, and simplicity*. Technical Report, Carnegie-Mellon University.

Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of data Analysis Techniques and Strategies*, *3*, 281-299.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

Shariff, A. A., Zaharim, A., & Sopian, K. (2009). The comparison of logit and probit regression analyses in estimating the strength of gear teeth. *European Journal of Scientific Research*, *4*, 548-553.